



HSE AI
AI and digital science
institute of FCS



fall into **MML** 2025

4th conference on
machine learning & AI

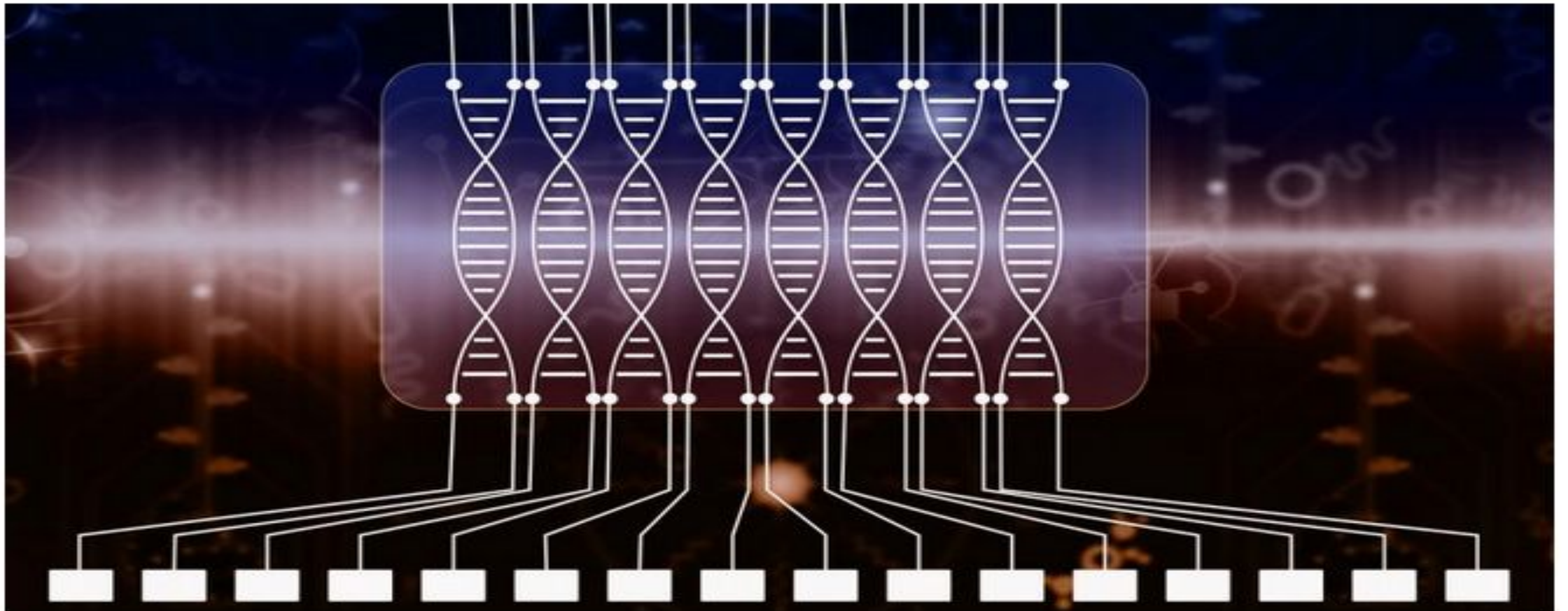
Maria Poptsova

24-25
october

Moscow, HSE University

OmicsFUSION: A GenAI model for Encoding,
Reconstruction, and Representation
of the Omics and DNA Data

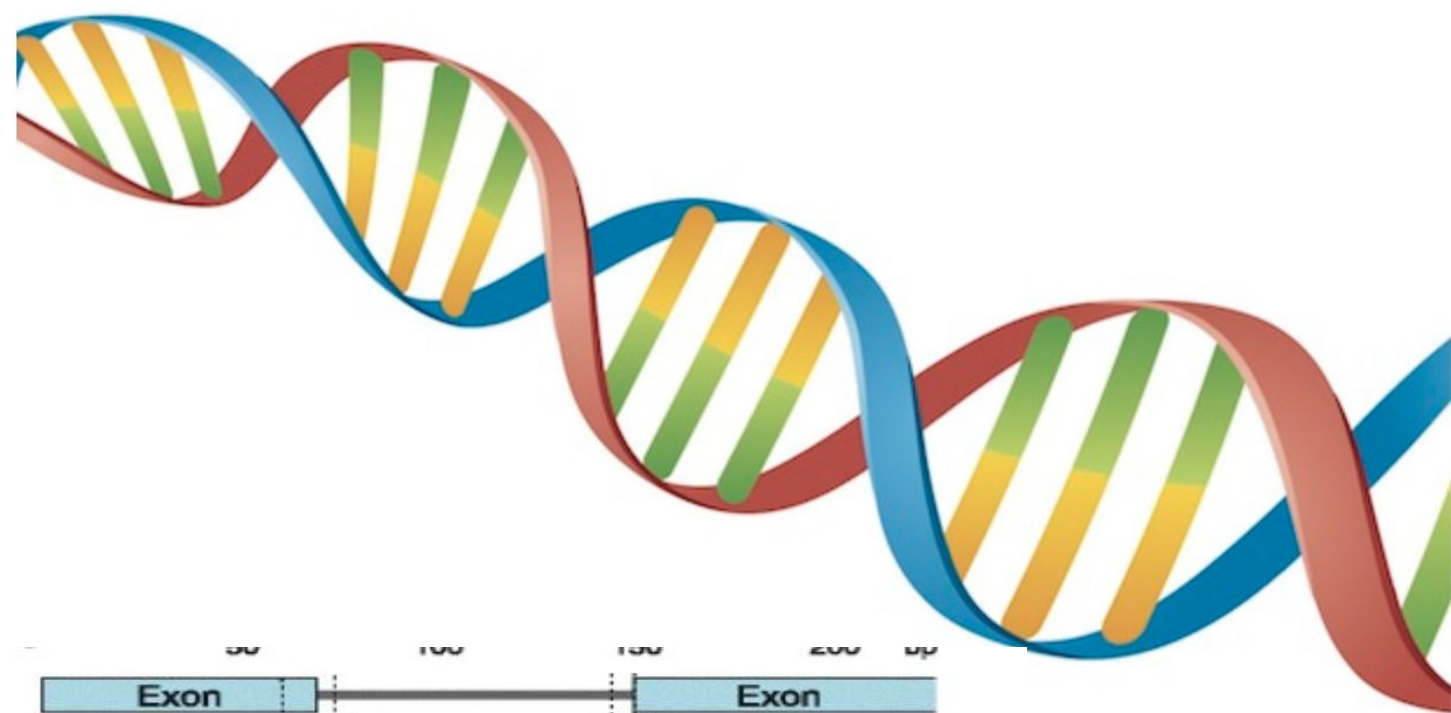
Algorithms are encoded in the genome



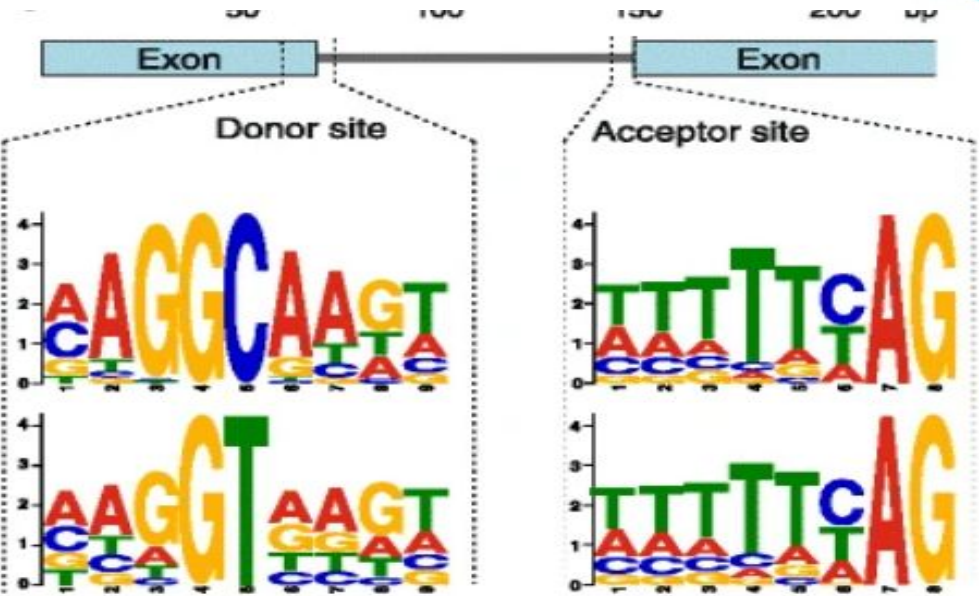
Layers of genomic encoding

- DNA sequence
 - Encoding into sequence: genes, exons, introns, promoters (some), enhancers (unclear), transcription factor binding sites

DNA-seq, RNA-seq, ChIP-seq



GATCTTAAACATTTATTCCCCCTGCAAACATTTTCAATCATTACATTGTCATTTCCCCTCCAAATTAAAT
TTAGCCAGAGGCGCACAAACATACGACCTCTAAAAAAGGTGCTGTAACATGTACCTATATGCAGCACCCT
ATATGAGAGCGGCATAGCAGTGTTTAGTCACTTGGTTGCTTTGTTTATATTAAGTGTGTTTTA
GCTATTGAGTTTAAACAAAGGGAGCGGTTTACATTGAATTAAAGGCAACTACTGATGGGTTGTGTAATGT
TTCAAAGAGCTGTTGCAGCATGAGTGGAAAATAAAACCGTATTAGTGCTGCCTGGCCCAGTTTGGCACAA
AATGGAGCGATTCCATTAAGAGAACGATTCAGCATAAGTGGAACAGCTAAAGTTTATGAAAATTTTAAAT
CTGGATGTAGAGAATCTCATAACACAGAAACAGCACTCCTAAAGATGCATTTATACTTCTGCATAGAGCA
CACAAGTATGCTTCAGCACAACTGTGCATGGTCACATAGCCCTTGCTGTGGCAGTCAGACAATGATGCA
CACCTCAATGAATAGAGCAAGCTGTGATTGGTCAGCTTTGTAC
CATCTTCAGTCAGTAATTTGGGGGCCATGTCTGACAGCAACCT
AACTCAATTTCTTCATCGTAATGTTGCCAAAATGACGAGCTATC
ATGCTTTCATTACCTCTCGGTTAGACTACTGTAATGAGTTACTA
CAAAGTGCAGCTGGTTCAAATGCAGCTGCTGGAGTGCTTTC
TCTTTCATCGGTGCATTGACTCCCATTAATATTTAATTTACAAT
AATGCTTTAGCTGCCAGTATTTGAGGGAGCTCCTGGTGTATT
ACCCTAAAGTGACCCACGAGTCTACATATTGATCCTGAAGGGT



	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA TER	UGA TER	A
	UUG Leu	UCG Ser	UAG TER	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCG Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Hidrophobic - Imino

Hidrophobic - Aromatic

Hidrophobic - Aliphatic

Polar - Acid

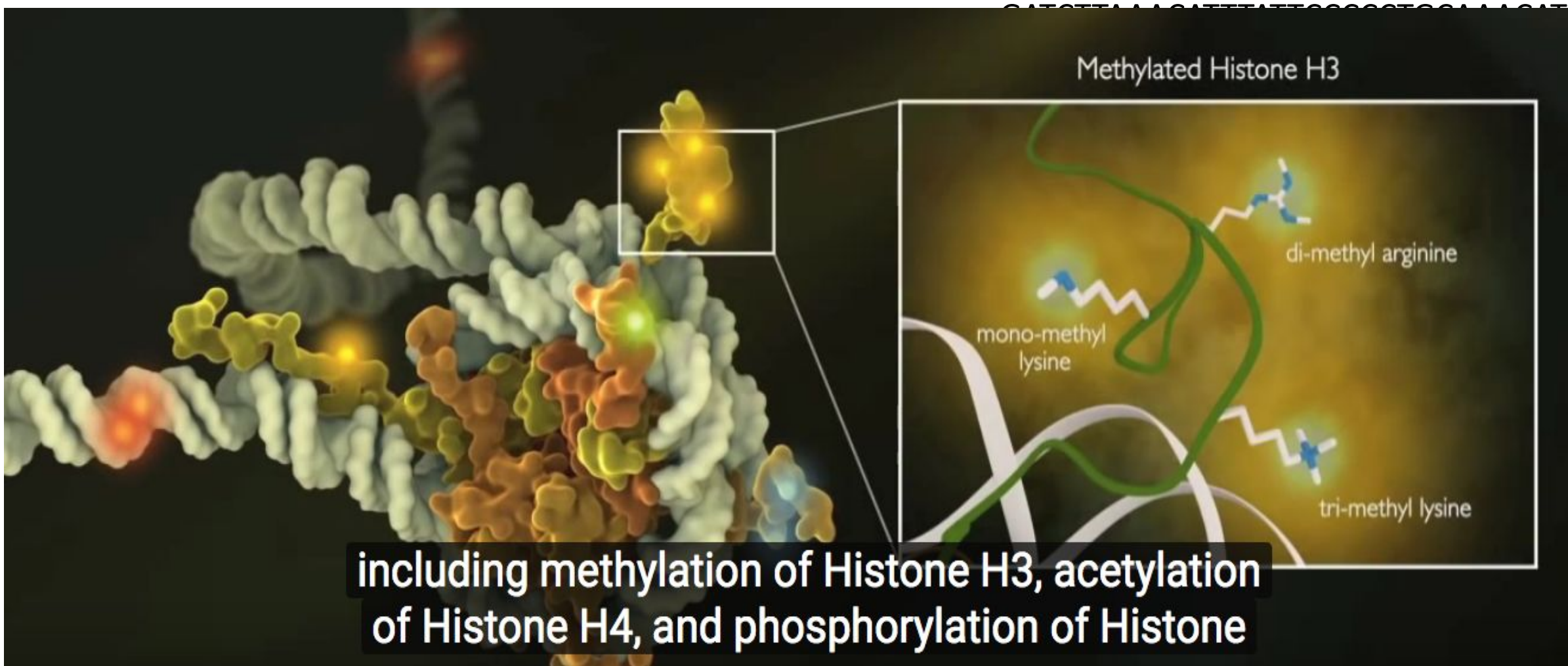
Polar - Neutral

Polar - Basic

Layers of genomic encoding

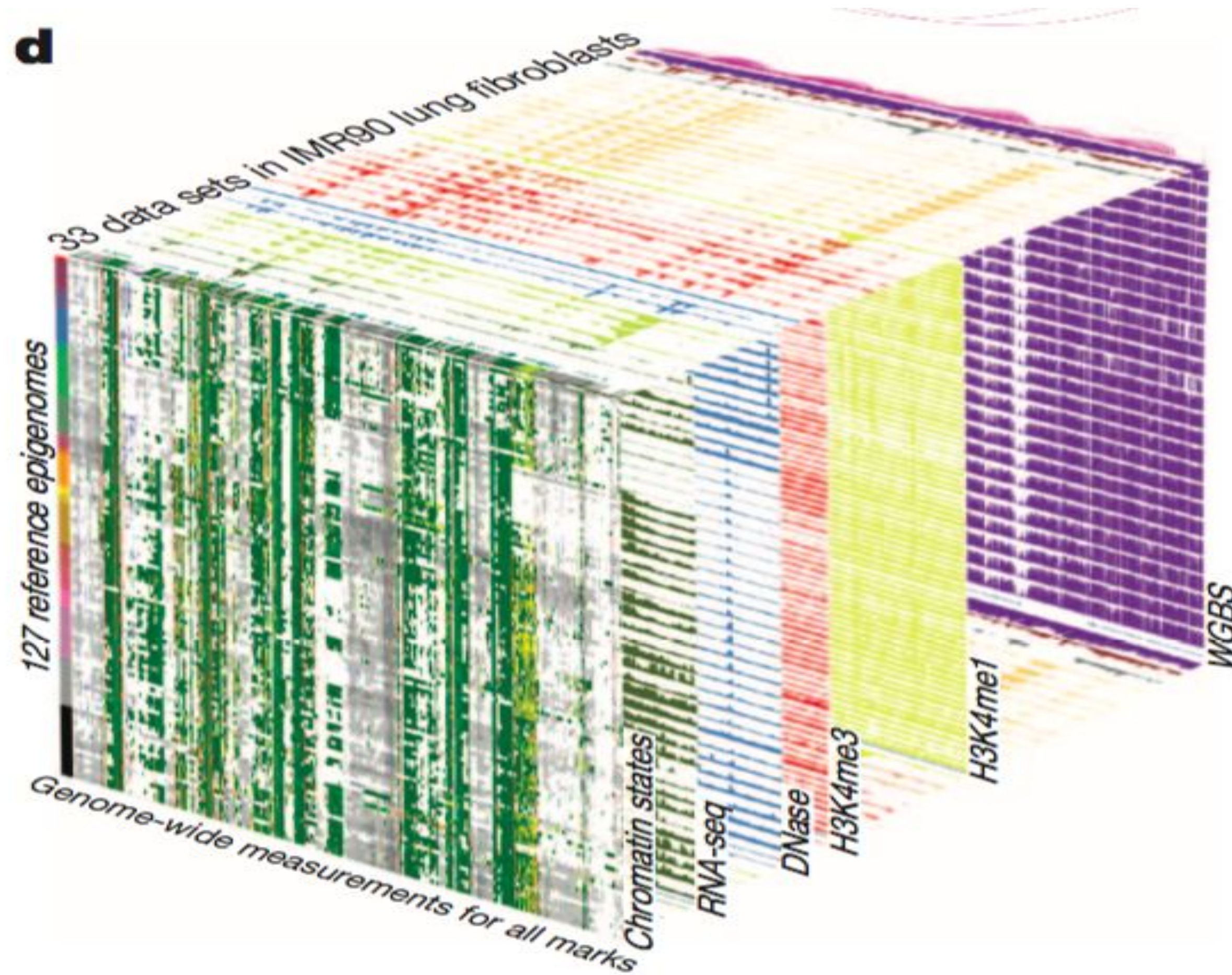
ChIP-seq, Bisulfite, MeDIP-seq,
MAD-seq, CLEVER-seq, Mnase-seq,
Hi-C, Hi-3C and more....

- DNA marks: Epigenomics
 - Methylations, histone modifications, chromatine packaging



GATCTTAAGATTATTCCTGCTGTAAGATTTCATTCATCATTACATTGTCATTTCCCCTCCAAATTAAAT
AAAAAAGGTGCTGTAACATGTACCTATATGCAGCACCCT
CTTGGTTGCTTTGTTTATATTAACCTTGAAAGTGTGTTTTA
TACATTGAATTAGGGCAACTACTGATGTTGTGTAAATGT
AATAAAACCGTATTAGTGCTGCCTGGCCAGTTTGGCACAA
CAGCATAAGTGCCACAGCTAAAGTTTATATAATTTTAAAT
CAGCACTCCTAAGATGATTATACTTTGCATAGAGCA
GGTCACATAGCCCTTGCTTGGCAGTCAGACAATGATGCA
GTCAGCTTTGTAGAGCTACGTGTGCCCTAACCTAGTCCT
CTGACAGCAACCTGTGTAAGGCCAGATGAAGCATCTGTAA
ATGACGAGCTATCAATAACTGATGCAGAAAAGCTTATTC
TAATGAGTTCTAGGCGGATGCCCTGCCGGCCTCACTAA
CTGGAGTGTTCTAGAACTAAGAAATCACAGTATTCTAGT
ATTTAATTTAATGTTAGTGTTTAAGTACAAAACCTC
CTCCTGGTTATTATAGAGCATGAGATTGATTCTTTAAG
GATCCTGAAGGGTCAGTCTGTACACCAGCCGGTACCTCT

Representation of all layers of genome encoding



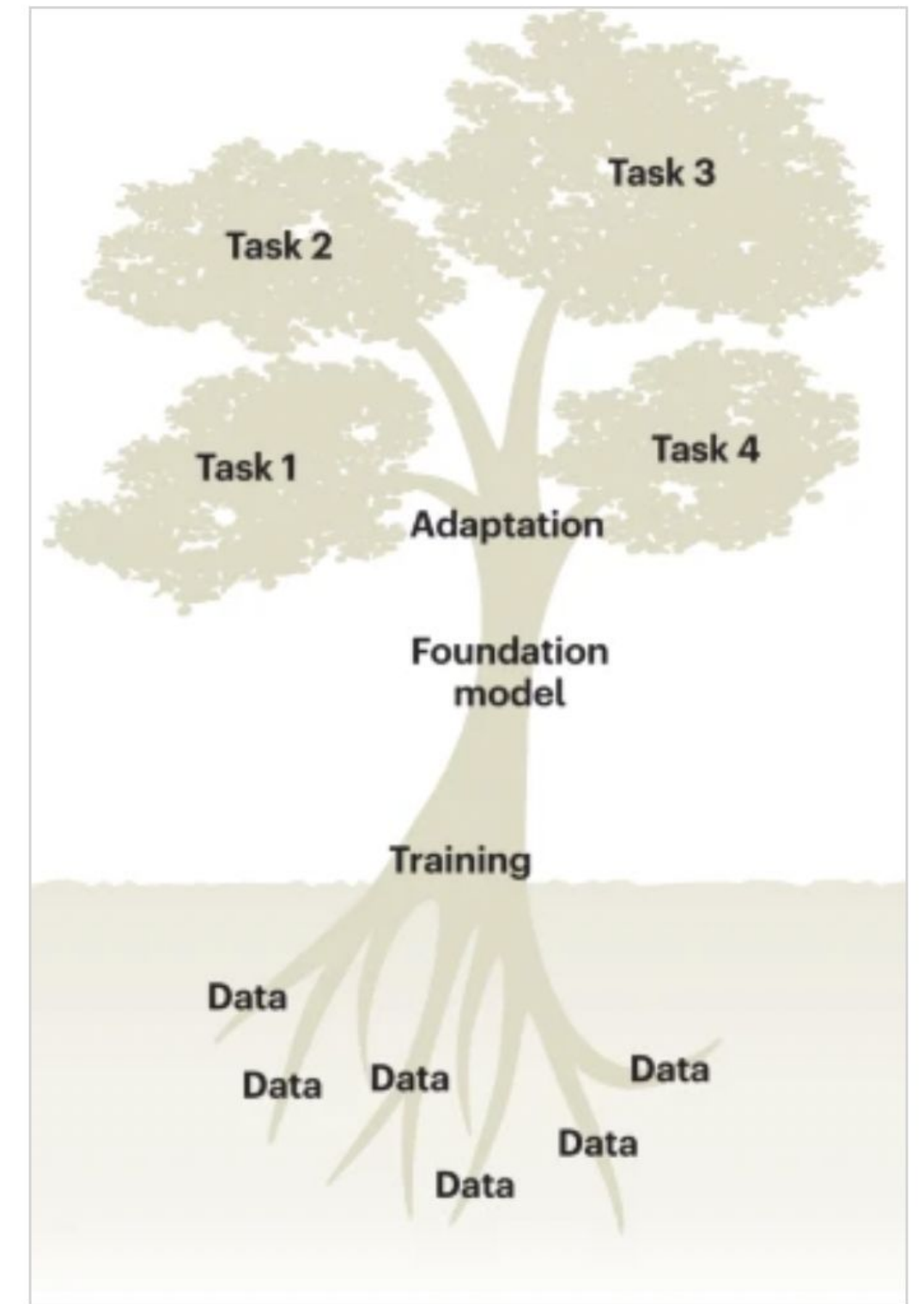
~ 100-200
ТЫС

~ 1-10
МЛН

Foundation models (FMs) in genomics

- to learn generalizable features from unstructured whole genome data
- fine-tune for a number of tasks
 - location and function of genes
 - regulatory elements
 - analyzing the evolution of specie

DNA sequences are orders of magnitudes longer (e.g. the human genome is 3.2 Billions of nucleotides) with long-range dependencies and interactions that span over 100K+ nucleotides in length



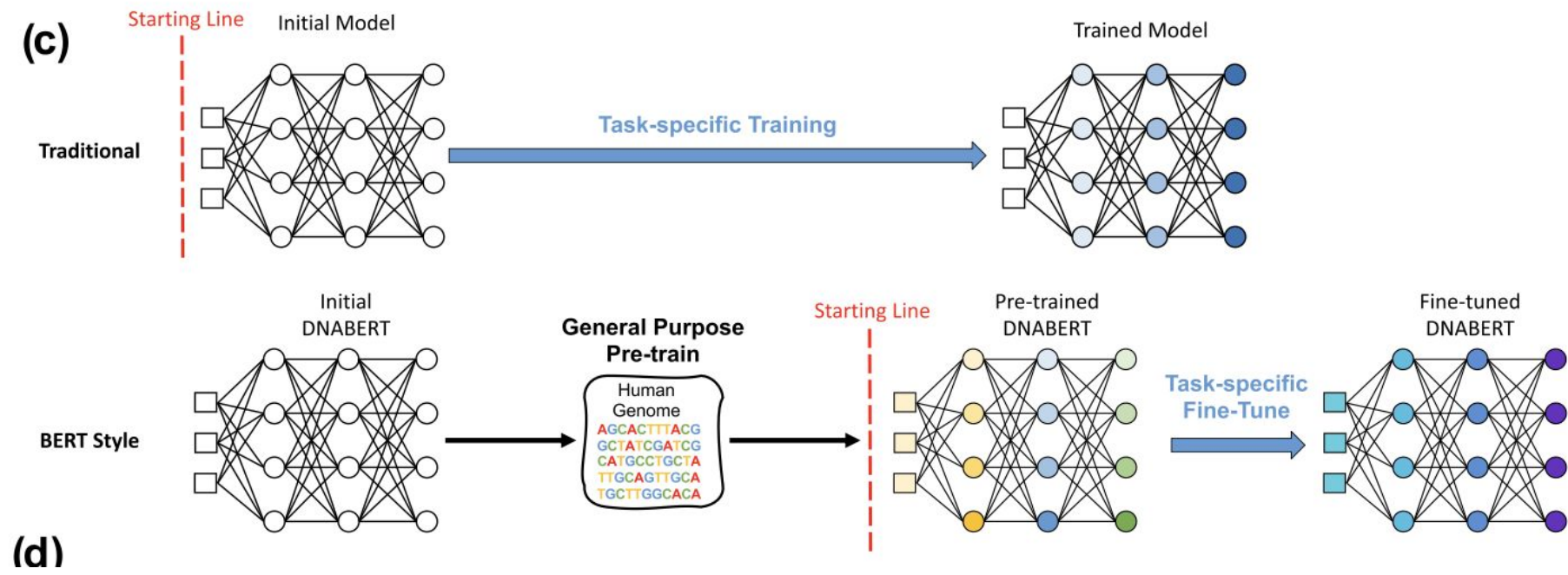
Conceptual illustration of a foundation model for multiple specific tasks. Credit: Debbie Maizels/Springer Nature

Genome analysis

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

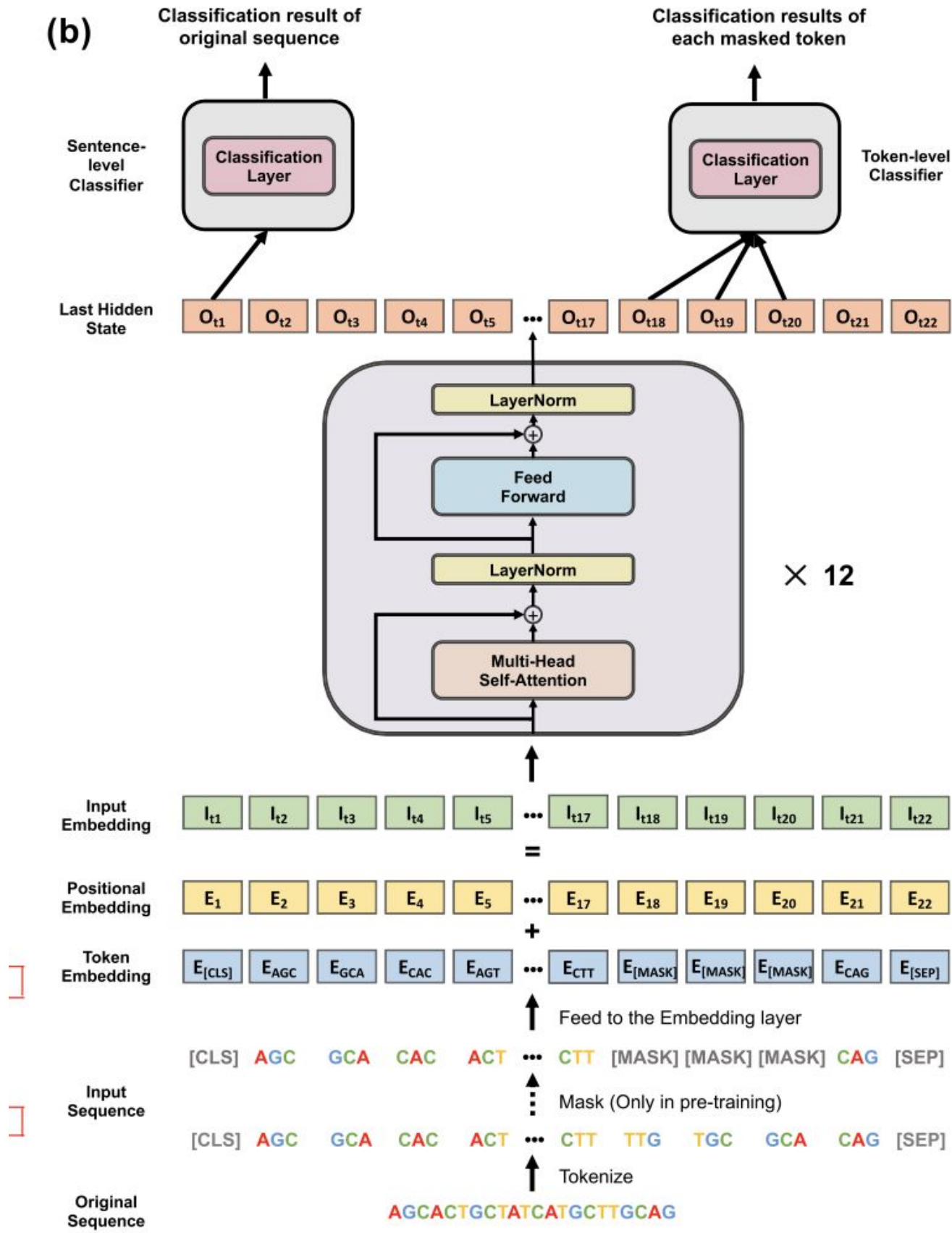
Yanrong Ji^{1,†}, Zhihan Zhou^{2,†}, Han Liu^{2,*} and Ramana V. Davuluri ^{3,*}

¹Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA, ²Department of Computer Science, Northwestern University, Evanston, IL 60208, USA and ³Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA



Pre-training: mask contiguous k-length spans of certain k-mers, (total ~15% of input sequence) - inferred from the immediately surrounding k-mers

Fine-tuning: no masking step, the tokenized sequence is directly fed to the Embedding layer.



DNABERT first takes a set of sequences represented as k-mer tokens as input



HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution

Eric Nguyen^{*,1}, Michael Poli^{*,1}, Marjan Faizi^{2,*},
Armin W. Thomas¹, Callum Birch Sykes³, Michael Wornow¹, Aman Patel¹,
Clayton Rabideau³, Stefano Massaroli⁴, Yoshua Bengio⁴, Stefano Ermon¹,
Stephen A. Baccus^{1,†}, Christopher Ré^{1,†}

arXiv

November 15, 2023

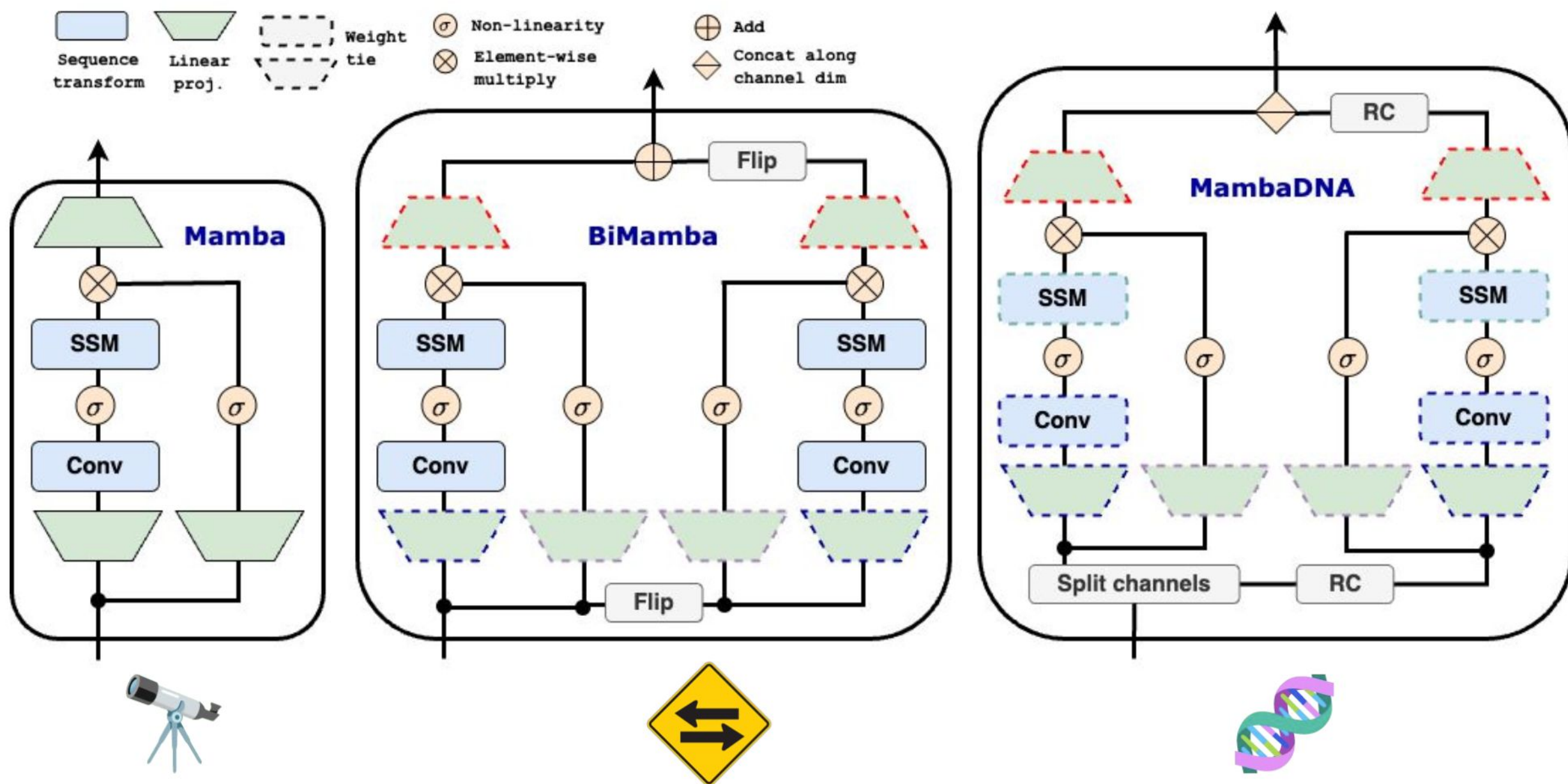
HyenaDNA, a **genomic foundation model** pretrained on the human reference genome with **context lengths of up to 1 million tokens at the single nucleotide-level** – an **up to 500x increase** over previous dense attention-based models.

HyenaDNA scales **sub-quadratically in sequence length** (training up to 160x faster than Transformer), uses single nucleotide tokens, and **has full global context at each layer**.

Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling

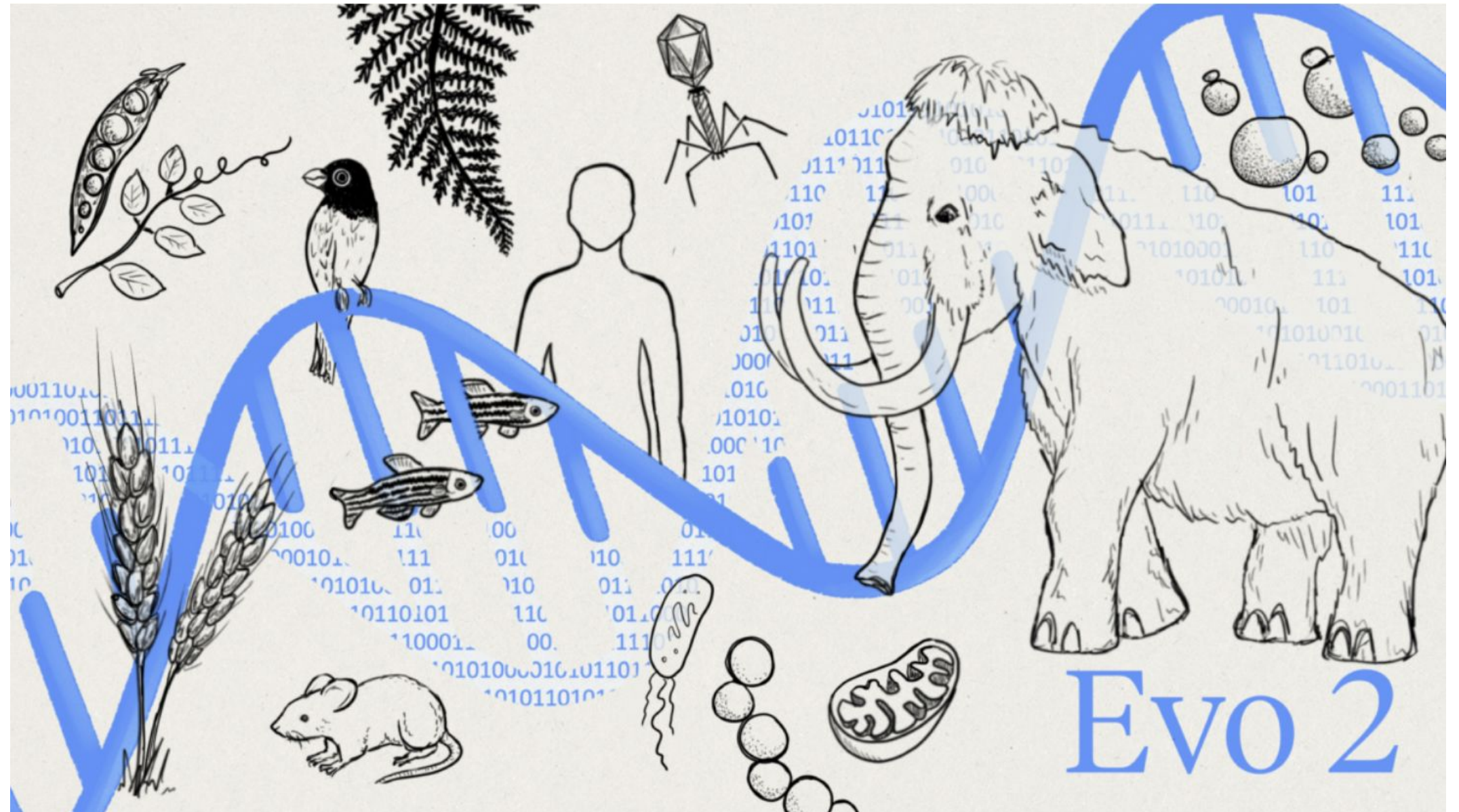
Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, Volodymyr Kuleshov

Building towards Caduceus



Evo-2 – самый большой “Chat-GPT” в геномике

Выучила
9,3 триллиона пар ДНК 128 000
организмов



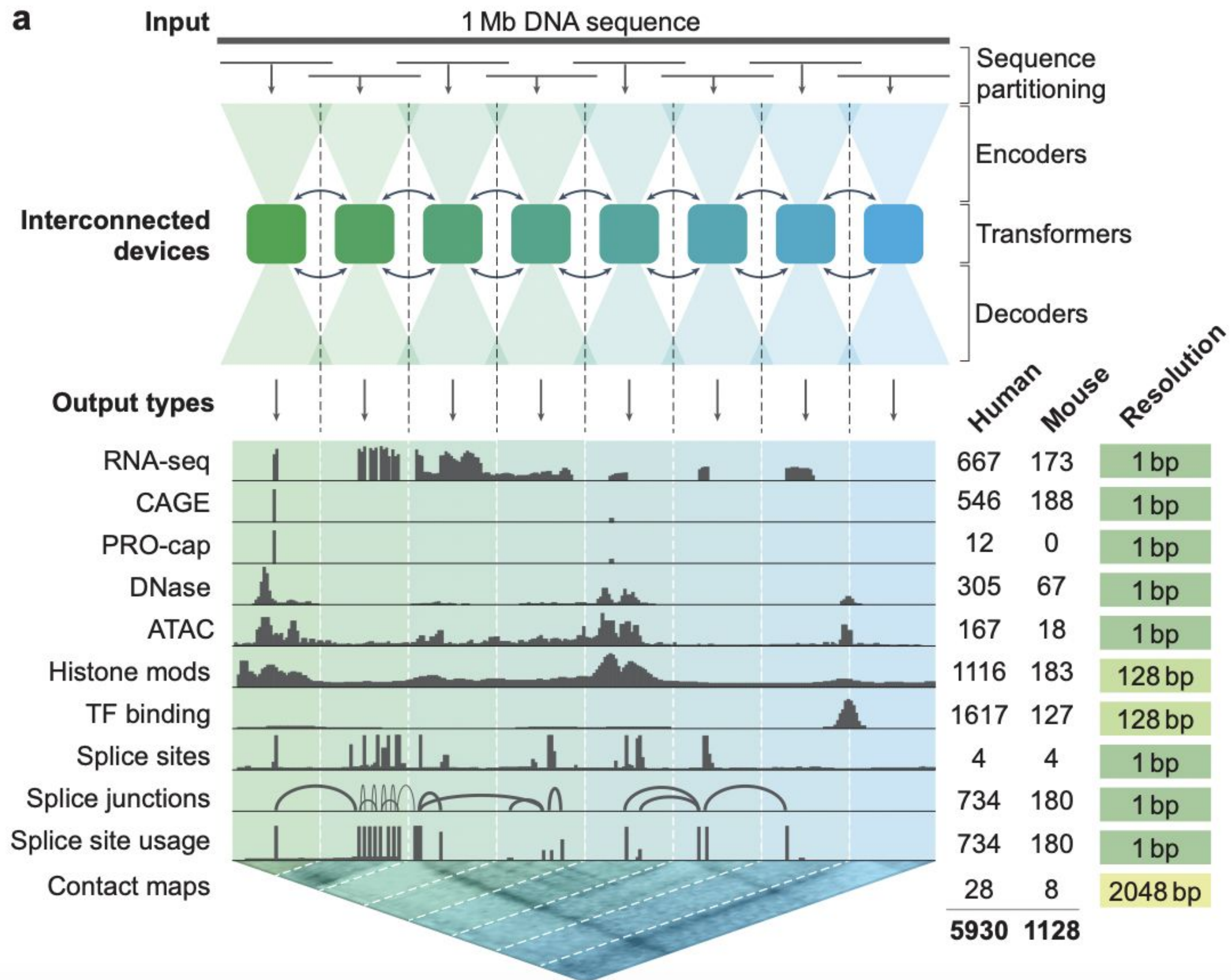
The largest Evo 2 model was trained using 2,048 NVIDIA H100 GPUs using NVIDIA DGX Cloud . As part of NVIDIA's partnership with Arc.

AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model

 Žiga Avsec,  Natasha Latysheva,  Jun Cheng,  Guido Novati,  Kyle R. Taylor,  Tom Ward,  Clare Bycroft,  Lauren Nicolaisen,  Eirini Arvaniti,  Joshua Pan,  Raina Thomas,  Vincent Dutordoir,  Matteo Perino,  Soham De,  Alexander Karollus,  Adam Gayoso,  Toby Sargeant,  Anne Mottram,  Lai Hong Wong,  Pavol Drotár,  Adam Kosiorek,  Andrew Senior, Richard Tanburn,  Taylor Applebaum,  Souradeep Basu,  Demis Hassabis,  Pushmeet Kohli

doi: <https://doi.org/10.1101/2025.06.25.661532>

AlphaGenome employs an encoder-decoder framework that integrates convolutional and transformer components, drawing inspiration from models such as Enformer² and Borzoi³. It introduces several key architectural features, including a U-Net-style decoder incorporating residual (skip) connections from the encoder. The model processes information through five core components: (1) a sequence encoder, (2) a transformer tower, (3) pairwise interaction blocks, (4) a sequence decoder, and (5) task-specific output heads. AlphaGenome has approximately 450 million trainable parameters (20% in the encoder, 28% in the sequence transformer, 15% in the pairwise blocks, 25% in the decoder, and 12% in the output embedding and prediction heads). The model was implemented using JAX⁷² and Haiku⁷³.

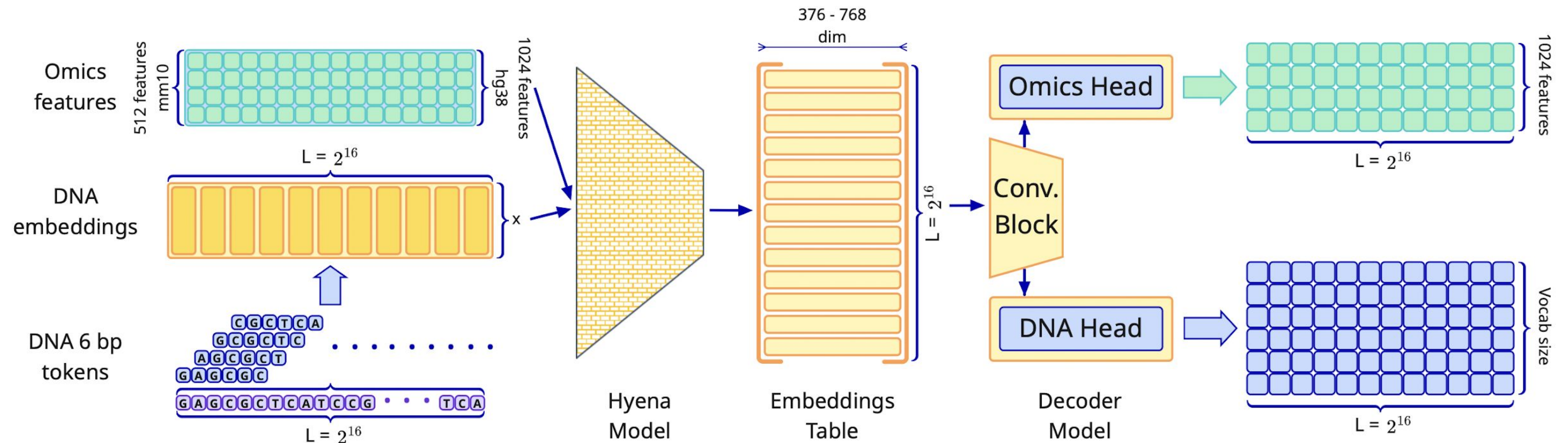


OmicsFUSION: A Pretrained Hyena-Based Framework for Encoding, Reconstruction, and Representation of the Omics and DNA Data

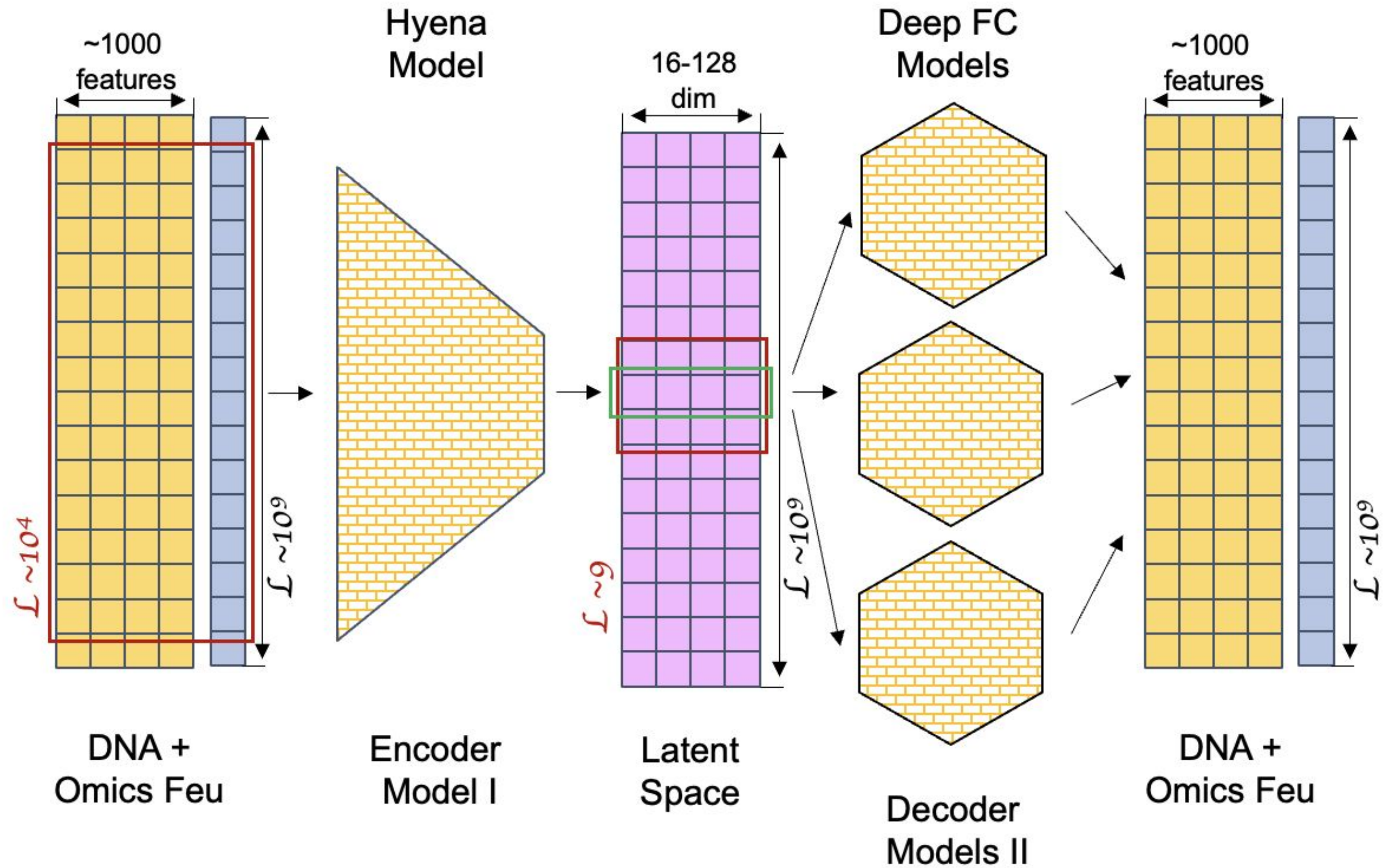
Nazar Beknazarov^{1*}, Nikita Pavlichenko¹, Artem Bashkatov¹, Maria Poptsova¹, Alan Herbert^{1,2}

¹International Bioinformatics Laboratory, HSE University, Moscow, Russia

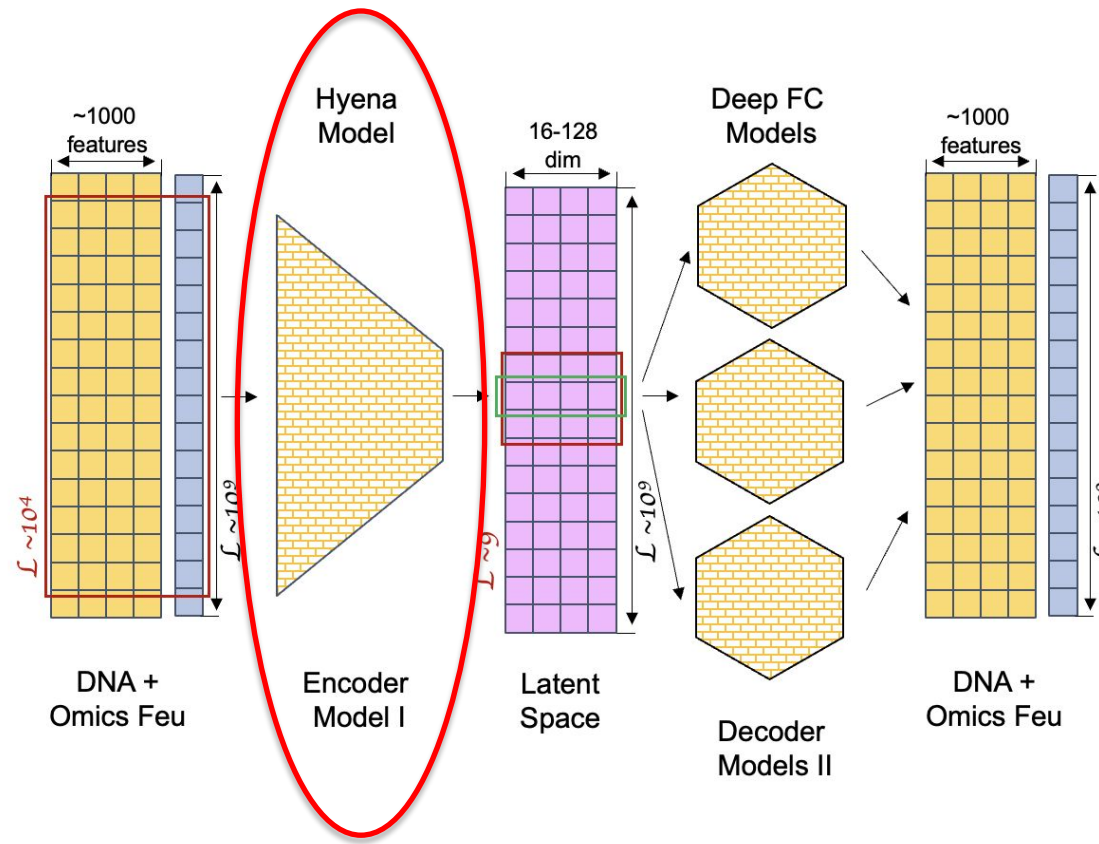
²InsideOutBio, Charlestown, MA, USA



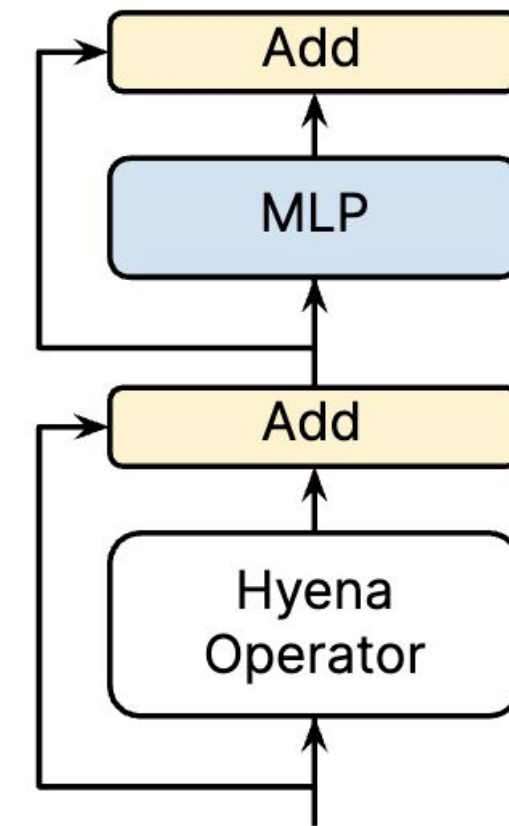
OMICS-DNA-Hyena



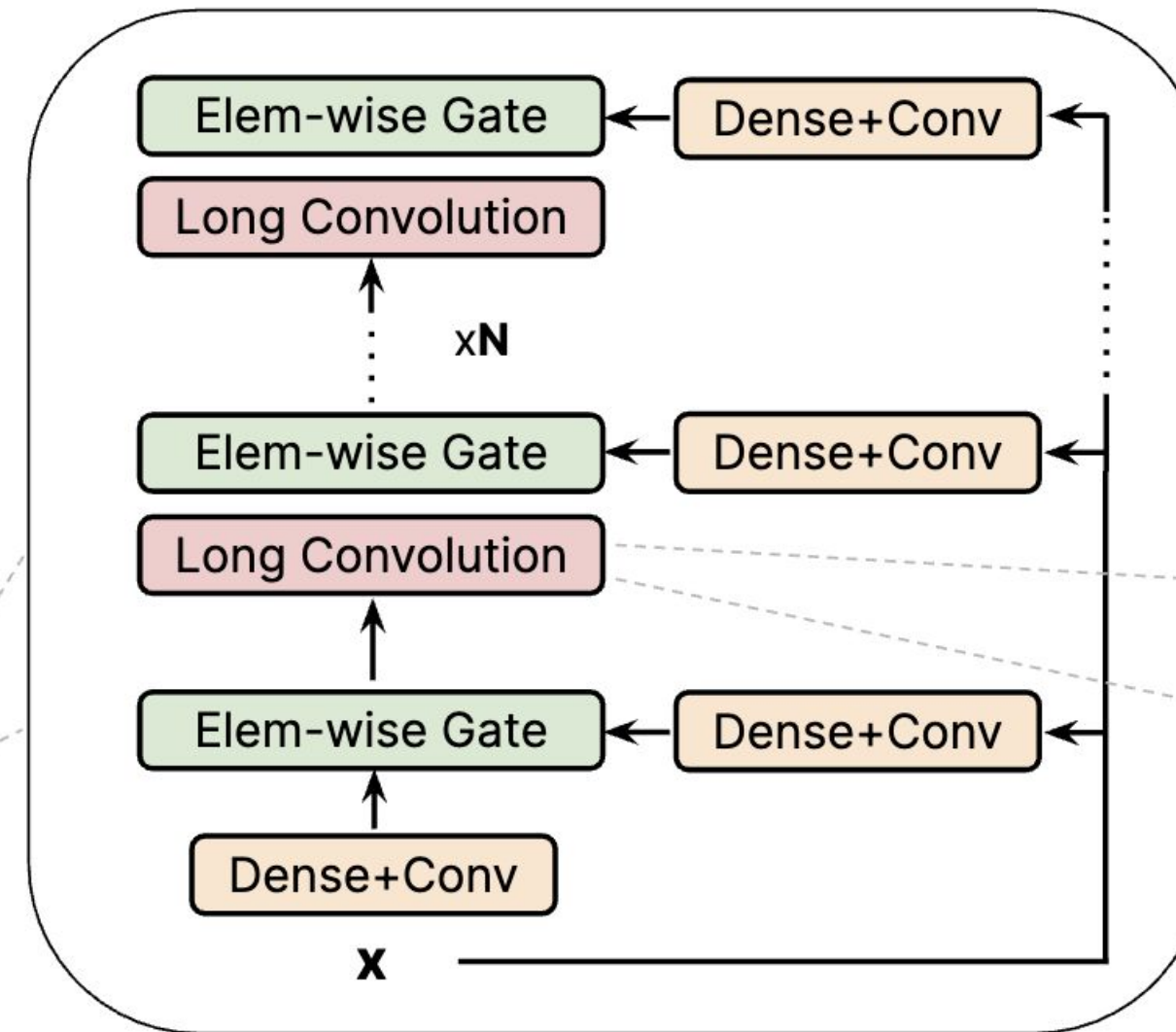
OMICS-DNA-Encoder



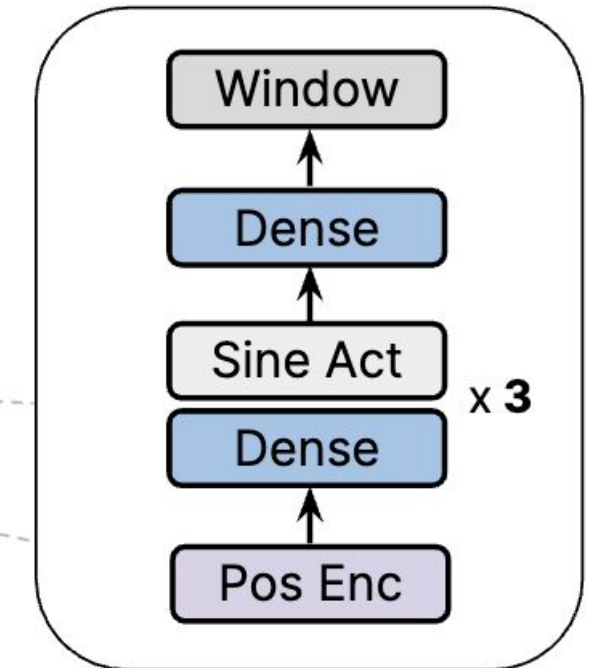
HyenaDNA Block



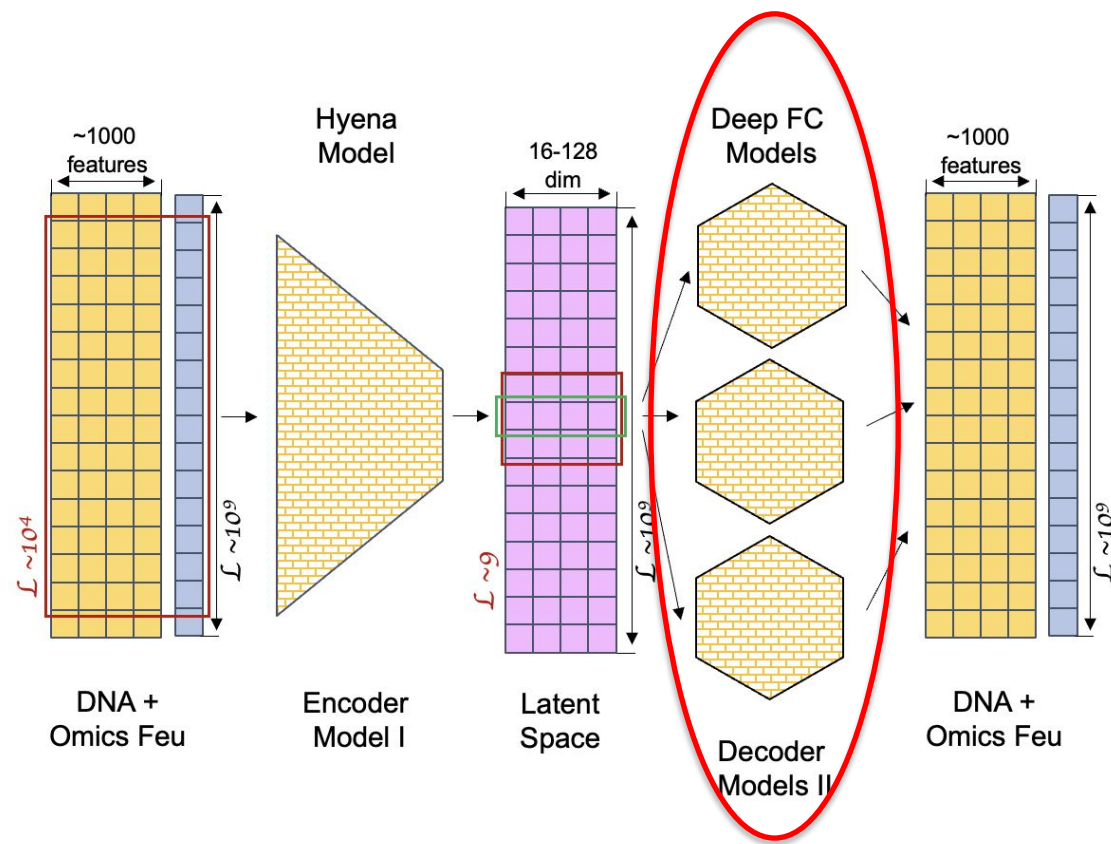
Hyena Order-N



Hyena Filters

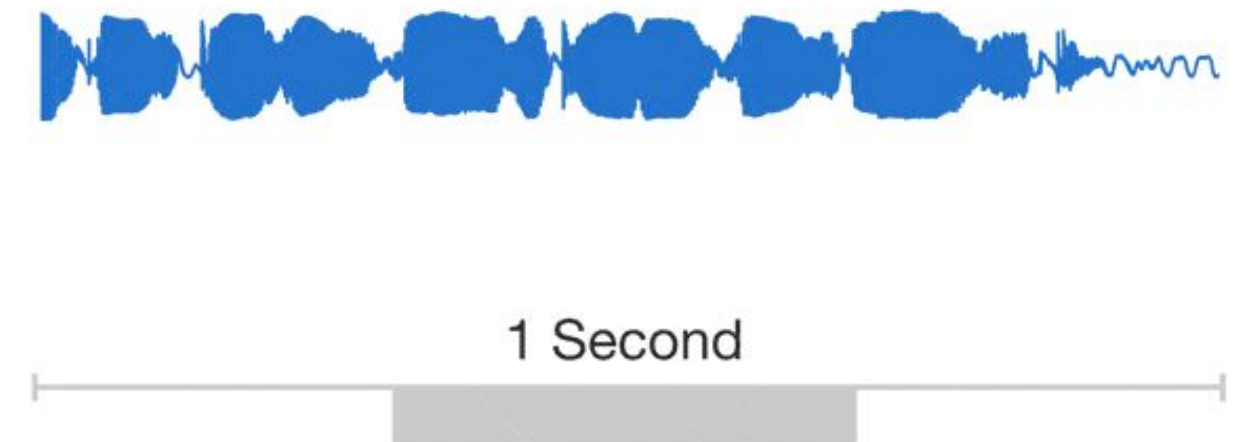


- Embedding layer
- 12x Hyena Layers



OMICS-DNA-Decoder

WaveNet



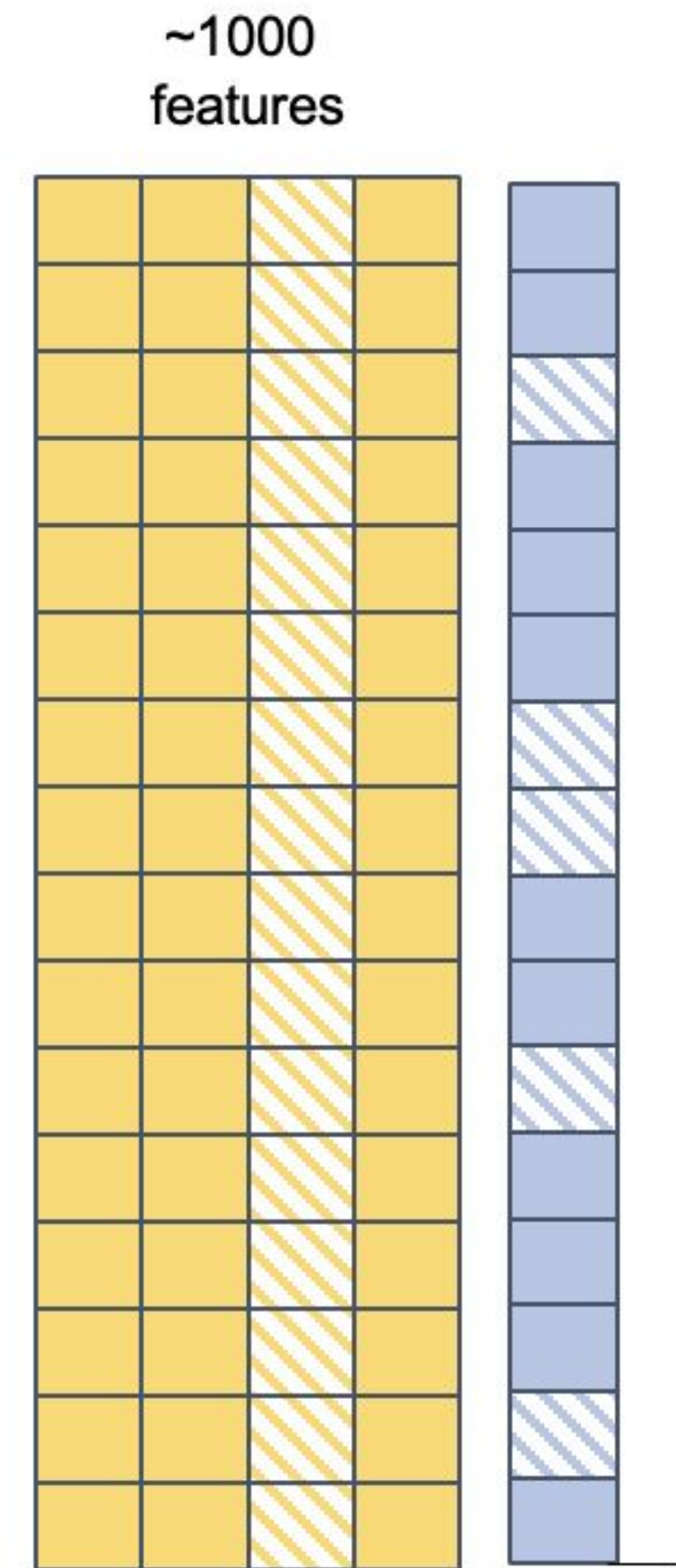
- CNN Block
- WaveNet Block
- Token Reconstruction Head
- Omics Reconstruction Head
- Sentence Reconstruction Head

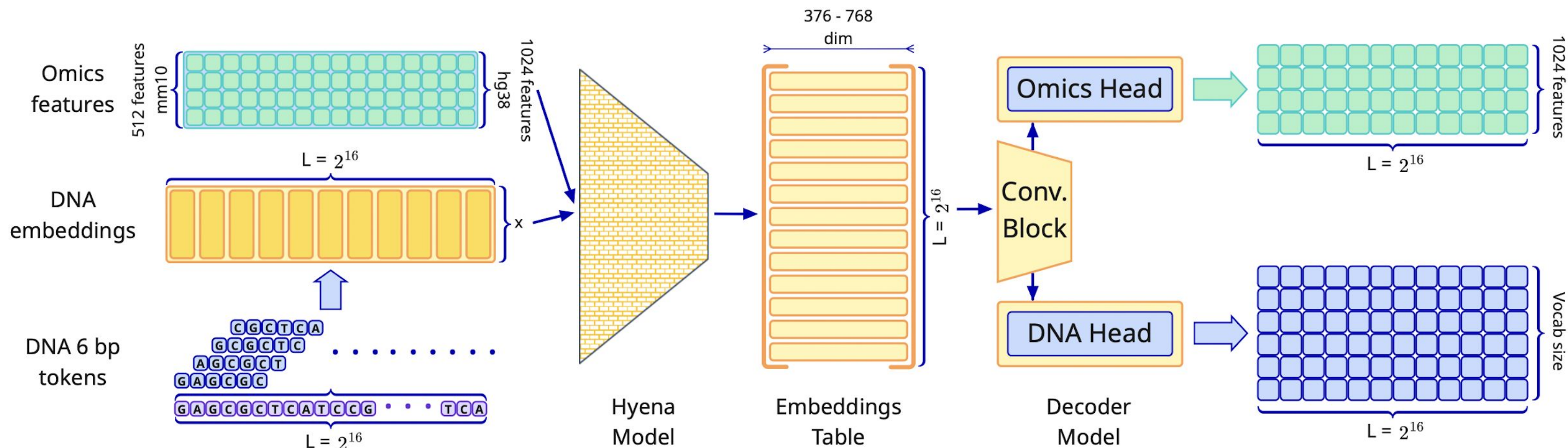
OMICS-DNA Masking

Task – to reconstruct data in autoregressive manner

To reconstruct two types of masking:

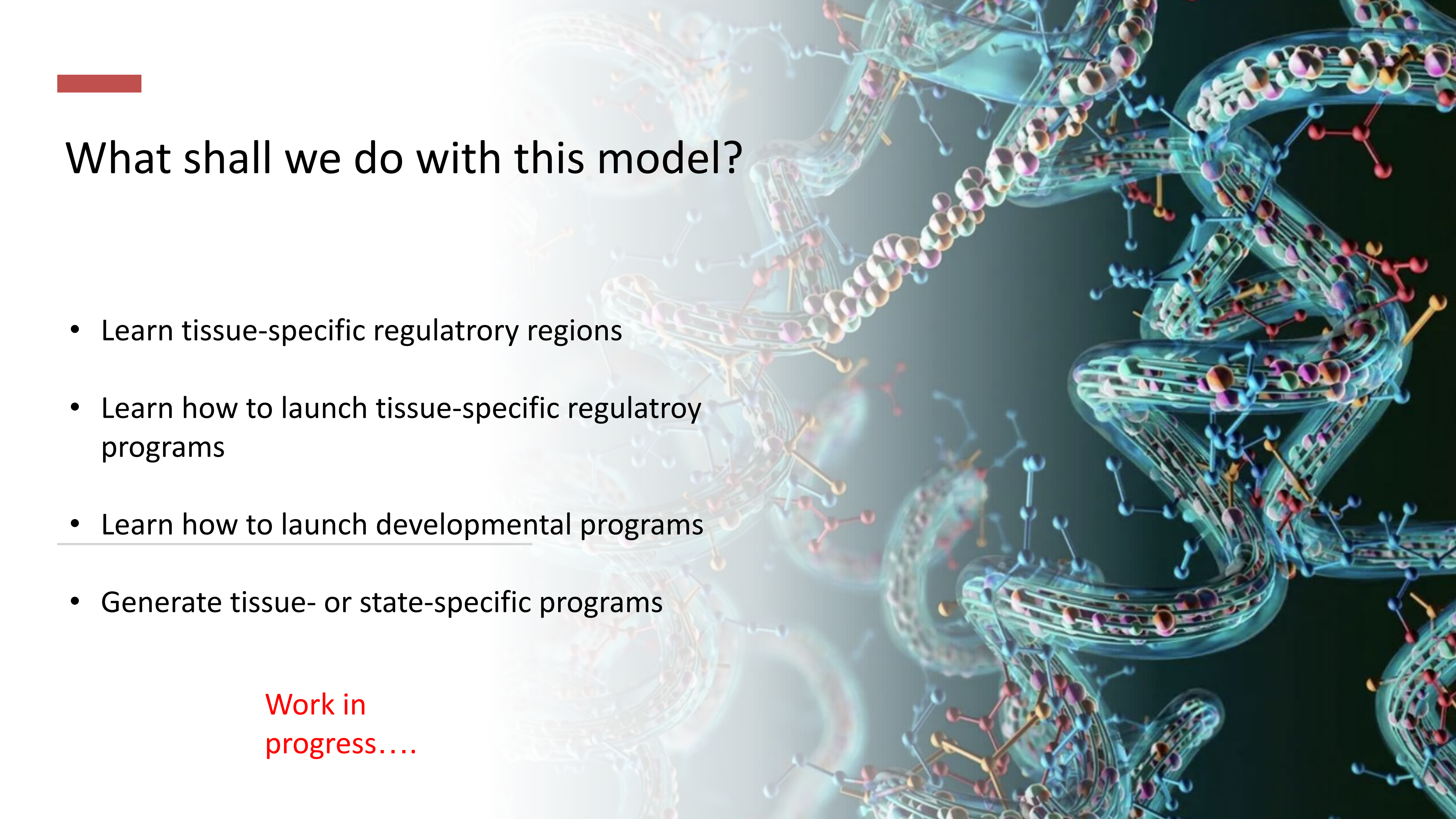
- Token masking
- Omics masking





Name	Non-Masked Token Acc.	Masked Token Acc.	Non-Masked Omics Acc.	Masked Omics Acc.	Non-Masked Omics F1	Masked Omics F1	Non-Masked Omics ROC AUC	Masked Omics ROC AUC
hg38_384	0.911	0.368	0.997	0.996	0.685	0.590	0.996	0.991
hg38_512	0.913	0.388	0.998	0.998	0.714	0.573	0.997	0.992
hg38_768	0.915	0.403	0.998	0.997	0.741	0.560	0.998	0.993
mm10_384	0.908	0.352	0.993	0.989	0.812	0.637	0.996	0.960
mm10_512	0.911	0.368	0.994	0.988	0.839	0.629	0.998	0.980
mm10_768	0.913	0.389	0.994	0.992	0.848	0.666	0.998	0.988

Table 3: Evaluation metrics for different models on hg38 and mm10 datasets. Accuracy, F1, and ROC AUC are reported for both masked and non-masked omics regions reconstruction and DNA token reconstruction tasks. All metrics are calculated on test set.



What shall we do with this model?

- Learn tissue-specific regulatory regions
- Learn how to launch tissue-specific regulatory programs
- Learn how to launch developmental programs
- Generate tissue- or state-specific programs

Work in
progress....

Международная лаборатория биоинформатики



Год создания – 2021

Миссия: Разработка и применение современных методов машинного обучения и искусственного интеллекта для биоинформатики и медицины



Попцова М., кандидат наук –
заведующая лабораторией



Герберт А., доктор наук –
научный консультант

26 сотрудников:

- 1 доктор наук
- 2 кандидата наук
- 2 PhD ученых до 35-ти лет
- 23 молодых ученых до 35-ти лет

Направления



Методы ИИ в биоинформатике



Кардиогенетика



Цифровой двойник клетки



Прогностические системы ИИ в
медицине

Текущие проекты

Разработка концепции
генетического
компьютера



Разработка
кардиогенетического теста с
использованием методов ИИ
«Кардиожизнь»

КАРДИО — **ИЗНЬ** —

Создание прогностических систем с
использованием методов ИИ для
предсказания рисков развития
кардиозаболеваний



✓ Магистерская программа «Анализ данных в биологии и медицине» аккредитована
Альянсом в сфере искусственного интеллекта

Партнеры



ЦЕНТРАЛЬНАЯ ГОСУДАРСТВЕННАЯ
МЕДИЦИНСКАЯ АКАДЕМИЯ
УПРАВЛЕНИЯ ДЕЛАМИ ПРЕЗИДЕНТА
РОССИЙСКОЙ ФЕДЕРАЦИИ

future inside
Genetico



Геномный
центр
РНИМУ



Бюджетное учреждение высшего
образования Ханты-Мансийского
автономного округа - Югры
«Сургутский
государственный
университет»



Городская клиническая
БОЛЬНИЦА №29
им. Н.Э.Баумана

100000Я



БУ ХМАО - ЮГРЫ ОКРУЖНОЙ КАРДИОЛОГИЧЕСКИЙ
ДИСПАНСЕР «ЦЕНТР ДИАГНОСТИКИ
И СЕРДЕЧНО-СОСУДИСТОЙ ХИРУРГИИ»



Genotek

inside out bio
science for living well

